

# Analisi Predittiva

CT0429  
Secondo appello

Gennaio, 2022

Cognome: \_\_\_\_\_ Nome: \_\_\_\_\_

Matricola: \_\_\_\_\_ Firma: \_\_\_\_\_

## ISTRUZIONI (DA LEGGERE ATTENTAMENTE).

Assicuratevi di aver scritto nome cognome e matricola sia qui che sul file Rmarkdown disponibile su Moodle. Il tempo a disposizione per completare tutto l'esame (la parte scritta e la parte su Moodle) è di **90 minuti**.

Nessuno studente può lasciare l'aula fino a che la docente non avrà verificato che tutti abbiano consegnato sia il compito scritto che il file Rmarkdown. Dopo la consegna attendete che la docente dia il permesso di lasciare l'aula.



**Question 1** (3 points)

Nella stima di un modello lineare semplice sono stati trovati i seguenti valori per  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$  e  $\widehat{Var}(\hat{\beta})$ :

```
fit <- lm(y ~ x, data = df)
coef(fit)
```

```
(Intercept)      x
          0.4      1.5
```

```
vcov(fit)
```

```
              (Intercept)      x
(Intercept)      0.70 0.25
x                 0.25 4.00
```

Inoltre si ha che:

```
dim(df)
```

```
[1] 10  2
```

(i) Si costruisca un intervallo di confidenza per il coefficiente di regressione  $\beta_1$

(ii) Si indichi se è possibile rifiutare l'ipotesi nulla  $H_0 : \beta_1 = 3$  (VS  $H_1 : \beta_1 \neq 3$ )

**Question 2** (6 points)

Si prenda in considerazione il dataset `df` le cui statistiche descrittive sono presentate in seguito:

```
summary(df)
```

x_continua	x_factor	y
Min. :0.01136	group1:31	Min. : 6.032
1st Qu.:0.23959	group2:28	1st Qu.: 36.865
Median :0.43692	group3:41	Median : 55.167
Mean :0.47617		Mean : 56.451
3rd Qu.:0.70468		3rd Qu.: 73.256
Max. :0.99803		Max. :105.952

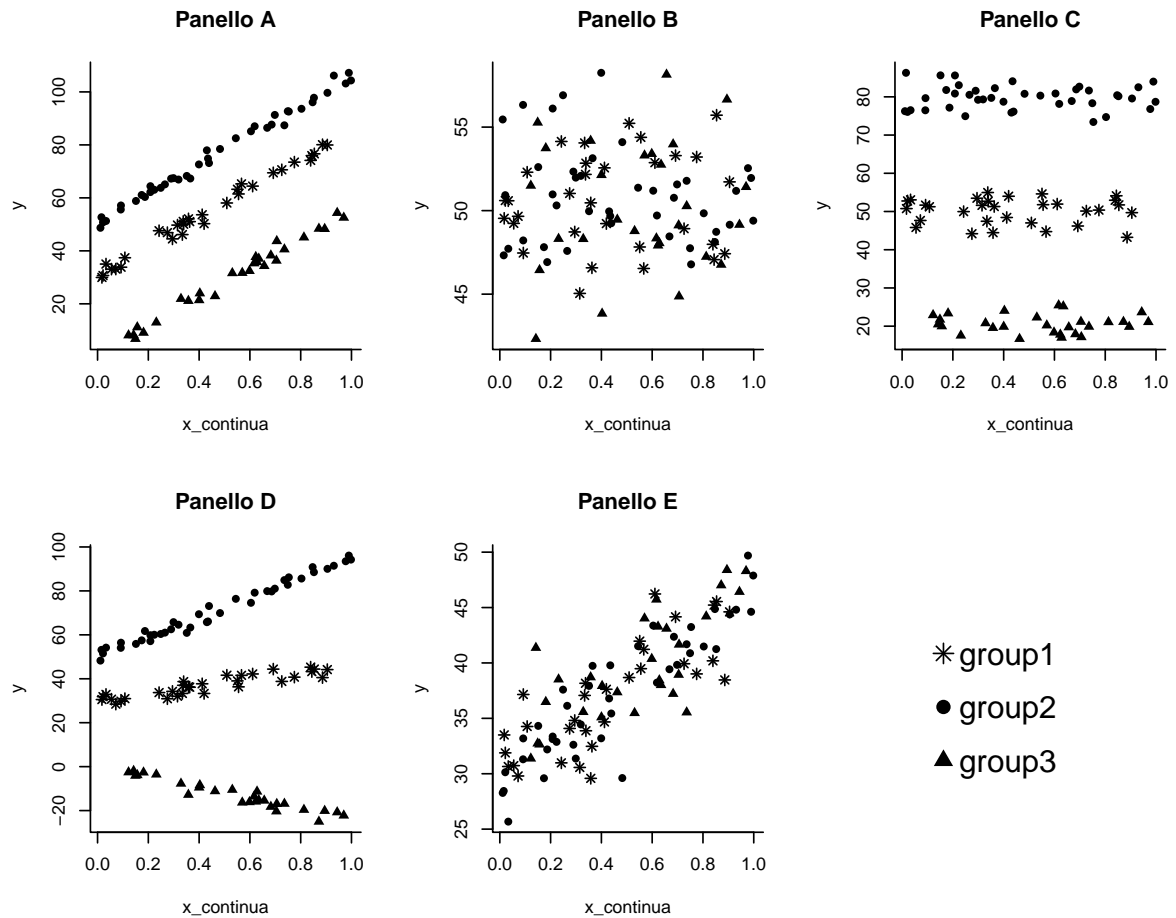
```
table(df$x_factor)
```

group1	group2	group3
31	28	41

Vengono stimati i seguenti modelli:

```
fitNull <- lm(y~1, data = df)
fitC <- lm(y~x_continua, data = df)
fitF <- lm(y~x_factor, data = df)
fitCF <- lm(y~x_continua+x_factor, data = df)
fitCFint <- lm(y~x_continua*x_factor, data = df)
```

Per ognuno dei modelli stimati si indichi il numero di gradi di libertà usati dal modello (cioè il numero di coefficienti di regressione stimati per ogni modello) e si indichi quale delle configurazioni dei dati nei pannelli nella figura a pagina successiva potrebbe essere meglio descritta da ognuno dei modelli. Infine si indichi per quale dei modelli si ottiene il maggior possibile valore di  $R^2$ .

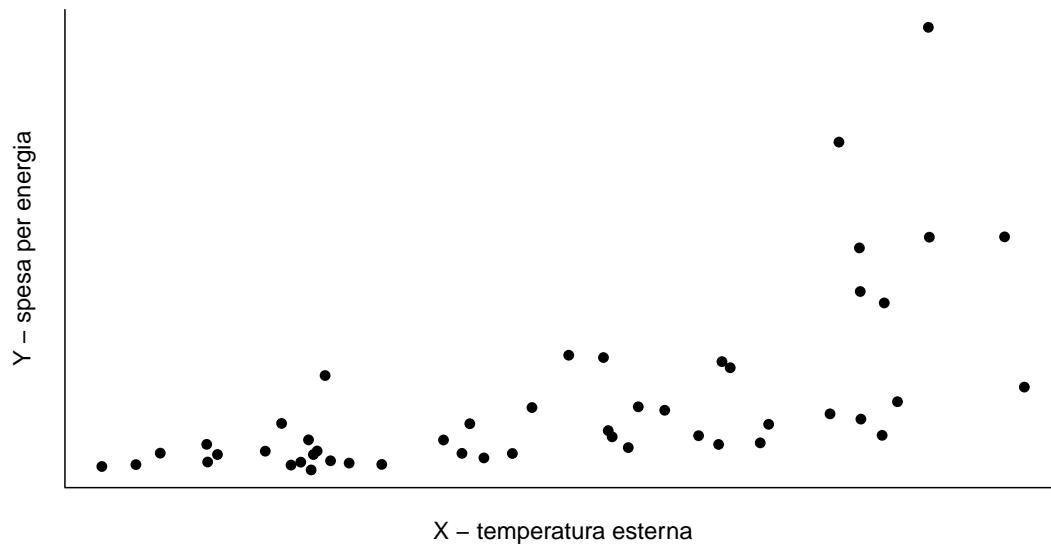


Modello	No. gradi libertà usati	Pannello
fitNull		
fitC		
fitF		
fitCF		
fitCFint		

- Il modello per cui si ottiene il maggior valore di  $R^2$  è \_\_\_\_\_

**Question 3** (3 points)

Una società che gestisce svariati data-centers desidera indagare la relazione tra temperatura esterna ed il costo legato alla spesa per mantenere le sale macchine entro determinati valori di temperature. La relazione tra le due variabili, misurata in un campione casuale di 49 giornate in diversi data-centers, è mostrata nel grafico sottostante:



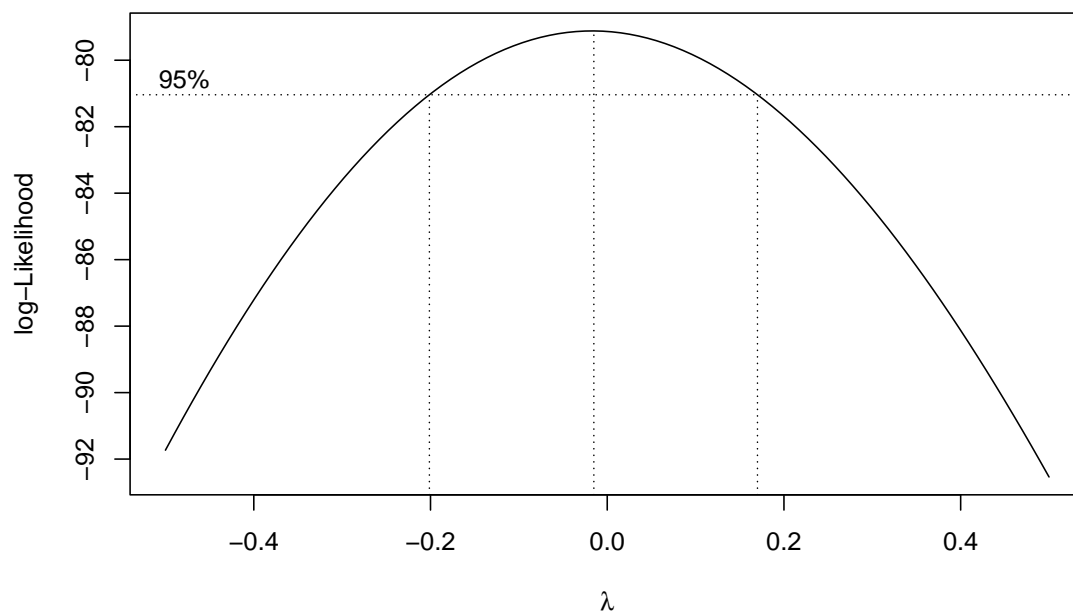
L'analista incaricato di analizzare i dati desidera usare un modello lineare semplice per costruire un primo modello predittivo, ma ritiene necessario, prima di procedere alla stima del modello, utilizzare una trasformazione di Box-Cox per trasformare la variabile risposta.

**Nota:** la trasformazione di Box Cox è definita come segue:

$$y_{\lambda} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

```
bctransf <- MASS::boxcox(y~x, data = df, lambda = seq(-0.5,0.5, by=0.05))
bctransf$x[which.max(bctransf$y)]
```

```
[1] -0.01515152
```

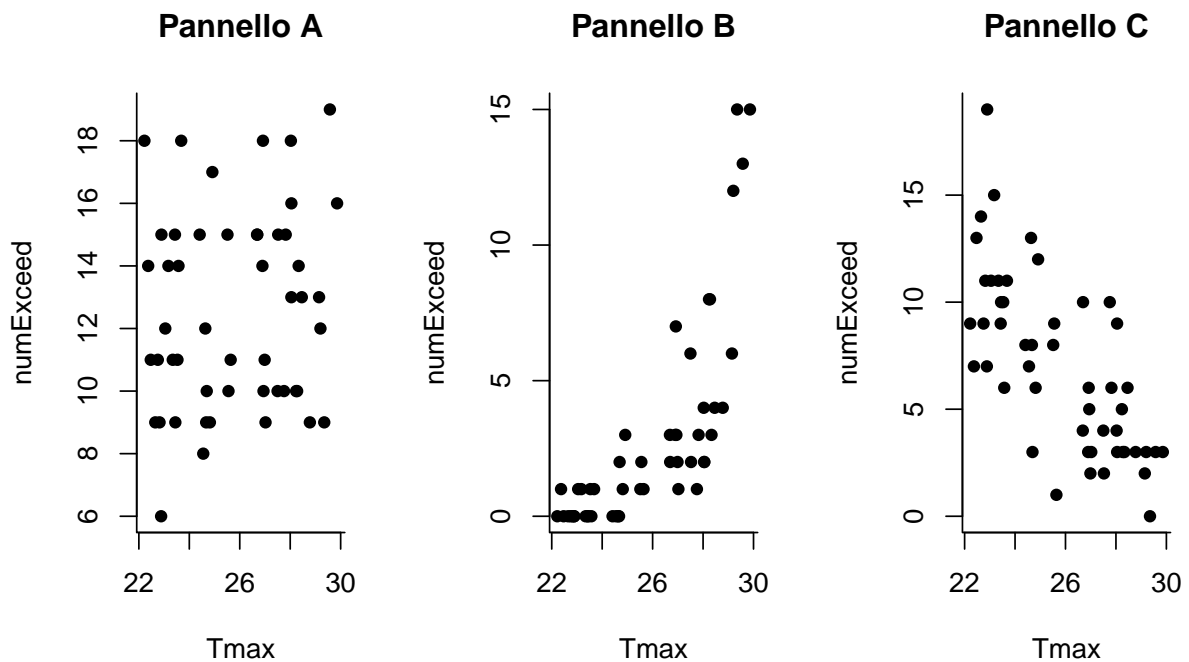


Si indichi se la scelta dell'analista di trasformare la variabile risposta sembra essere ragionevole, esplicitando le possibili motivazioni a favore o contro la scelta dell'analista.

Si indichi poi cosa si può evincere dall'output ottenuto dalla funzione `MASS::boxcox`: quale è la trasformazione che potrebbe venire applicata alla variabile risposta?

**Question 4** (3 points)

Una società che gestisce svariati data-centers monitora il numero di volte in una giornata in cui la temperatura nel data center eccede per più di 10 minuti un determinato limite. Si desidera indagare il possibile effetto della temperature massima registrata all'esterno del data-center ( $T_{max}$ ) sul numero di eccedenze del limite (`numExceed`). Le informazioni sulle due variabili sono registrate in un campione di 50 giornate estive in diversi data-center. I dati sono mostrati nella Figura sottostante e vengono fornite alcune statistiche riassuntive:



```
summary(df)
```

	Tmax		numExceed
Min.	:22.22	Min.	: 0.00
1st Qu.:	:23.47	1st Qu.:	: 0.00
Median :	:25.60	Median :	: 1.50
Mean :	:25.80	Mean :	: 2.88
3rd Qu.:	:27.98	3rd Qu.:	: 3.00
Max.	:29.86	Max.	:15.00

Per indagare la relazione tra temperatura e il numero di eccedenze viene usato un modello lineare generalizzato con una distribuzione Poisson per la variabile risposta e la funzione legame canonica:

```
fit <- glm(numExceed~Tmax, data = df, family = poisson())
summary(fit)
```

Call:

```
glm(formula = numExceed ~ Tmax, family = poisson(), data = df)
```



Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8386	-0.9016	-0.5114	0.7526	2.2815

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-14.2600	1.5382	-9.271	<2e-16 ***
Tmax	0.5646	0.0546	10.341	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 213.182 on 49 degrees of freedom  
Residual deviance: 50.331 on 48 degrees of freedom  
AIC: 157.24

Number of Fisher Scoring iterations: 5

- (i) Si indichi quale dei pannelli nella Figura nella pagina precedente è più probabile mostri i dati usati del dataset `df` usato per stimare il modello. Pannello \_\_\_\_\_
- (ii) Si scriva in forma esplicita il modello che viene stimato usando la funzione `glm`:

**Question 5** (3 points)

Una società che gestisce svariati data-centers desidera predire in quali giornate è possibile che la temperatura massima nella sala dei data-center ecceda un determinato livello considerato a rischio. Vengono quindi raccolte per un campione di 73 giorni e data-centers svariate informazioni su caratteristiche meteorologiche all'esterno del data-center e l'informazione binaria se la temperatura massima ha superato il livello pre-determinato.

```
table(df$highTemp)

0  1
30 43

fit1 <- glm(highTemp ~ Tmean, data = df, family = binomial())
coef(fit1)

(Intercept)      Tmean
      -3.0         0.2

deviance(fit1)

[1] 80.31498
```

Si usa poi il modello `fit1` per stimare la probabilità che in due giornate venga ecceduto il limite che indica una alta temperatura nella sala:

```
nd <- data.frame(Tmean = c(20,25))
rownames(nd) <- c("giorno 1", "giorno 2"); nd

      Tmean
giorno 1   20
giorno 2   25

predict(fit1, newdata = nd)

[1] NA NA
```

Infine si costruisce un modello in cui tutti i predittori sono usati:

```
fitAll <- glm(highTemp ~ ., data = df, family = binomial())
coef(fitAll)

(Intercept)      Tmean      wind  relHumid  sunshineHrs
   -0.8100    0.2300  -0.1700   -0.0067   -0.1200

deviance(fitAll)

[1] NA
```

- (i) Come si può interpretare il valore del coefficiente  $\beta_{T_{mean}}$  per il modello `fit1`?
  
  
  
  
  
  
  
  
  
  
- (ii) Usando il modello `fit1` si fornisca una stima della probabilità che nel “giorno 1” la temperatura massima ecceda il limite di temperatura predefinito (si indichi cioè il primo valore mancante dell’output di `predict`)
  
  
  
  
  
  
  
  
  
  
- (iii) Si indichi se la probabilità che la temperatura massima ecceda il limite di temperatura predefinito è più alta per il “giorno 1” or il “giorno 2” (usando il modello `fit1`)
  
  
  
  
  
  
  
  
  
  
- (iv) Si indichi se è possibile sapere quale dei modelli tra `fit1` e `fitAll` ha devianza maggiore. In caso affermativo si indichi quale dei modelli ha devianza maggiore.